近期阅读的文献加以总结，整理了下自己的思路。拟选择微博数据作为分析对象，挖掘特征数据，借助于语义网技术高效组织数据，以多种输入（关键字、节点图）方式改善交互，可视化地展示分析结果。

|  |  |
| --- | --- |
| 数据 | 文本（微博） |
| 问题 | 大规模数据分析效率不足 |
| 分析方法 | 数据挖掘+语义网技术+可视化 |
| 具体过程 | 见图1 |



图 1 微博数据可视化分析流程

在微博数据分析过程中，首先选择已经构建好的本体库（如FOAF或自定义），通过数据挖掘的方法（Topic Model）抽取微博中的关键词，将用户按语义网模型组织。根据高层的Tag，将用户分类，并以SOM图表示。

由于数据是以语义网模型组织，具有明显的层次结构，所以用户可以选择SOM图中局部内容，以LoD方式浏览数据，当浏览的数据达到一定程度后直接以节点图显示全部。在此过程中，记录用户浏览路径，方便退回。

同时，提供关键字搜索，结果以详细信息列表方式展示。

对于复杂搜索，用户输入完整句子，系统将自动建立语义节点图，减少用户交互。同时提供可以修改节点图的接口，让用户选择适当的Tag或连接边，解决词语歧义问题。建立好的节点图按照子图搜索在组织好的语义网数据中查找。

在此过程中，主要创新：

1. 以语义网模型组织微博数据。现有的语义网研究大都是针对已经组织好的语义网数据如FOAF、DBLP进行分析，数据源有DBpidea，Freepidea。然而对于微博数据分析手段主要还是借助于数据挖掘的方式。这里我们将尝试以语义网模型组织微博数据，并提出相应的分析方法。
2. LoD的SOM数据浏览方法。大尺度的数据集增加了可视化的难度，人们很难知道数据的整体结构。SOM适合于大数据集的可视化，特别是体现分类数据相互关系的情形。语义网模型先天的层次结构，能够很方便地进行LoD显示。
3. 灵活的输入。自动构建简单语义网能够简化用户的交互，由于语义网本身就是图结构，子图搜索则是很直接的技术。修改自动生成的语义网，可以借助用户的知识纠正语义分析过程中一词多义或多词同义的情况，提高搜索的准确性。

目前的困难：

1. 构建本体库。虽然存在一些本体库模型如FOAF, SIOC, DBLP，但是不一定满足分析目标。例如FOAF模型是描述个人信息和个人社会网络的本体模型，可以查找兴趣相投的人。但是对于某个人发表的信息无法很好地表达，因而难以追踪话题的发展形态。如果是寻找热点话题，只能另外寻找本体库模型。
2. 实现。由于整个过程步骤较多，而每一步都涉及到以前从未使用过的技术，实现起来需要花费很多时间。而论文截止期即将到来，这是一个极大的考验。如果去掉其中的部分环节能否作为一个完整的工作？

而作为长期目标，可视化分析平台可以做如下架构，如图2。

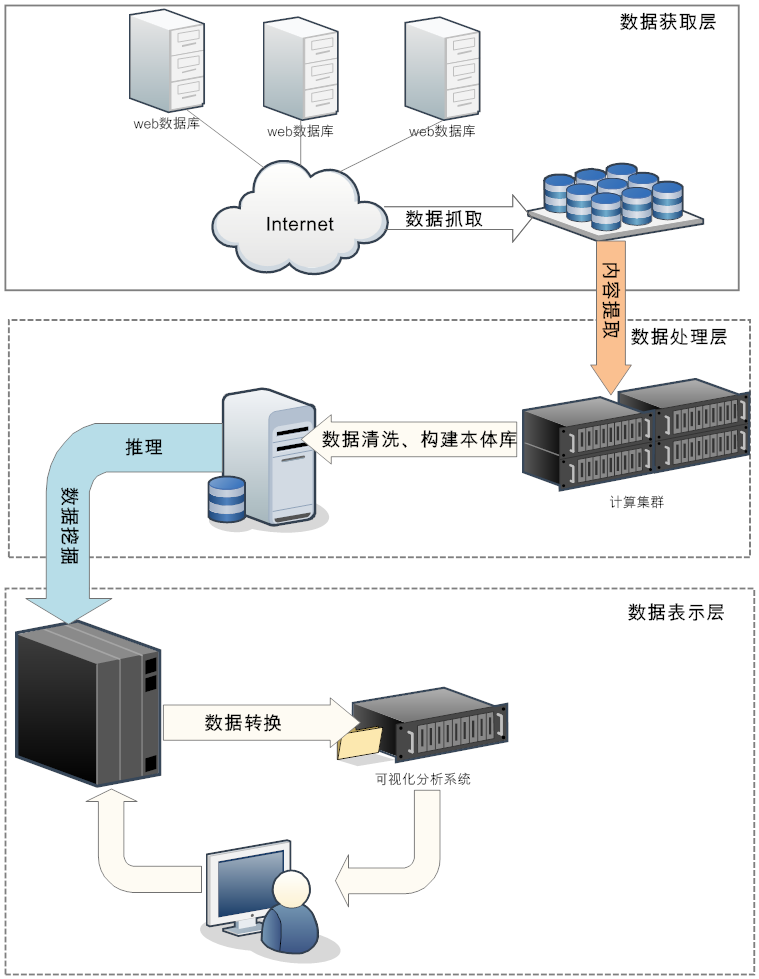


图 2 文本可视化分析架构

架构共分三个层次：

1. 数据获取层：负责从网络数据库中抓取原始数据，并将海量数据存入分布式数据库中。由于网络数据存在多源、异构、动态变化等特点，需要发现各类数据源，能够对异构数据进行存储，同时适应数据的动态变化。
2. 数据处理层：负责从存储的数据中提取适当的内容，对多源数据加以融合，以相同的数据结构表示异构数据。进行初步处理之后，需要对同一结构的数据加以清洗，通过本体库将数据组织成语义网模型，获得精确数据。
3. 数据表示层：负责数据的可视分析。深入挖掘语义网模型数据中的特征，将其转换成适当的可视化模型，经过绘制后展现给用户。用户交互式地浏览、搜索、过滤精确数据，从而发现隐藏在数据中的知识。